# Anmerkungen zur Faktorenanalyse

Blasius, Jörg

# Anmerkungen zur Faktorenanalyse

*Jörg Blasius**

**Abstract:** The article gives an overview on the method of factor analysis as a statistical tool for data reduction within the context of historical social research. Problems of the interpretation of latent variables and other parameters, factor rotation, and criteria for the extraction of factors are discussed.

## 1. Einleitung

Faktorenanalyse ist ein Oberbegriff für eine Vielzahl von Techniken zur Strukturierung von multivariaten Daten. In dieser Arbeit werden die wichtigsten Verfahren vorgestellt, sowie Möglichkeiten und Probleme bei deren Anwendung diskutiert. Die grundlegende Annahme bei der Faktorenanalyse ist, daß verschiedene Meßoperationen, die einen gemeinsamen Kontext bilden, auf eine dritte, nicht direkt meßbare (latente) Größe zurückzuführen sind. Das Ziel ist, diese Größe, die üblicherweise als Faktor, Dimension oder Achse bezeichnet wird, zu isolieren und mit Hilfe von beobachtbaren (manifesten) Variablen zu beschreiben.

Die ersten Anfänge der Faktorenanalyse reichen ins vorige Jahrhundert zurück. So war Galton (1869) insbesondere an der Klassifikation und Determination von Typen interessiert. Er und ebenso Spencer diskutierten die Existenz von allgemeinen und speziellen Fähigkeiten. Erste weitergehende Überlegungen stammen von Pearson (1901) und Spearman (1904). Spearman entwickelte die »2 - Faktorentheorie« für die Messung von Intelligenz, dem klassischen Beispiel für die Verwendung der Faktorenanalyse in der empirischen Sozialforschung. Seiner Theorie zufolge

* Address all communications to: Jörg Blasius, Universität zu Köln, Zentralarchiv für Empirische Sozialforschung, Bachemer Str. 40, D-5000 Köln 41.

gibt es einen allgemeinen und einen testspezifischen Faktor, die unkorreliert zueinander sind. Die erste algebraische Darstellung der Hauptkomponentenanalyse - dieses Verfahren ist meistens gemeint, wenn von Faktorenanalyse gesprochen wird - gab Hotelling (1933), der ein auf m Faktoren verallgemeinertes Modell vorstellte.

Die von Hotelling als »principal components« bezeichneten Hauptkomponenten lassen sich nicht direkt erheben, sondern lediglich mit Hilfe geeigneter Indikatoren beschreiben. Die Auswahl der Indikatoren sollte auf jeden Fall theoriegeleitet erfolgen, da es sonst zu inhaltlichen Artefakten kommen kann. Mit der Hauptkomponentenanalyse kann geprüft werden, welche Variablen den gleichen Kontext messen, d.h. mit welcher Hauptkomponente (Dimension, Achse) sie korrelieren (diese Korrelation von Variablen mit den Achsen wird in der Faktorenanalyse als »laden« bzw. »Ladung« bezeichnet) und wieviele dieser Dimensionen existieren.

Bei praktischen Anwendungen, wie z.B. der Auswertung von Intelligenztests, soll aus einer großen Anzahl von Variablen (z.B. Fragen zum mathematischen und künstlerischen Verständnis) herausgefunden werden, ob eine Person eher mathematische, naturwissenschaftliche, sprachliche oder künstlerische Fähigkeiten besitzt. Ferner wird dabei überprüft, welche Variablen sich als Indikatoren einer gemeinsamen Dimension erweisen, im Fall des Intelligenztestes mit welchen Variablen welche Fähigkeiten gemessen werden. Allgemein ausgedrückt: Mit der Faktorenanalyse soll eine Vielzahl von Variablen auf wenige Faktoren reduziert werden, mit denen dann eine Beschreibung der »Realität« möglich ist. Dies Prinzip der Datenreduktion gilt auch für andere multivariate Verfahren.

## 2. Anwendungsbeispiele

Bei der Faktorenanalyse gibt es sowohl explorative als auch konfirmatorische Verfahren. Sind keine oder nur vage Vermutungen darüber vorhanden, welche Variablen zusammenhängen und wie die entsprechenden Faktoren korrelieren, wird von explorativer Analyse gesprochen - dies heißt jedoch nicht, daß die Auswahl der Variablen willkürlich erfolgen kann. Können Hypothesen über den Zusammenhang von Variablen und über die Korrelationen der Faktoren angegeben werden, so können diese Hypothesen mit Hilfe der konfirmatorischen Faktorenanalyse getestet werden. Da beide Vorgehensweisen einander nicht ausschließen, ist es nicht unüblich mit Hilfe der explorativen Faktorenanalyse herauszufinden, welche Variablen eine gemeinsame Dimension beschreiben, daraus ein Kausalmodell abzuleiten und dieses dann mit der konfirmatorischen Faktorenanalyse anhand eines weiteren Datensatzes zu testen.

Für alle Verfahren der Faktorenanalyse gilt, daß auf keinen Fall beliebig viele Variablen gleichzeitig in die Analyse eingehen sollten, in der Hoffnung, daß sich schon irgendwelche Dimensionen ableiten lassen werden. Das trifft zwar fast immer zu, doch können auch solche Variablen auf einen gemeinsamen Faktor laden, die lediglich durch Zufall hoch miteinander korrelieren, ohne jedoch einen gemeinsamen inhaltlichen Bezug zu haben. Diese Nonsenskorrelationen führen u.U. zu einem weiteren Fehler: »Echte« vorhandene Strukturen werden nicht entdeckt, da diese durch stärkere »unechte« Strukturen überlagert werden. Generell gilt für alle Arten der Faktorenanalyse die Bemerkung von Mulaik (1986, S. 24), der darauf hinweist, daß mit blinder Exploration unintelligente Ergebnisse produziert werden.

Best (1986) verwendet die Hauptkomponentenanalyse zur Untersuchung des Abstimmungsverhaltens von Parlamentariern aus Deutschland und Frankreich in den Jahren 1848/49. Als Eingabeinformation benutzt er die Korrelationsmatrizen von 98 (Frankfurter Nationalversammlung) und 86 (Assemblee Nationale Constituante) Abstimmungen (ja/nein) in dem genannten Zeitraum. Damit untersucht er, ob es bei den Abgeordneten bestimmte Abstimmungsmuster gibt. Als ein Ergebnis erhält Best ein deutliches Links - Rechts - Schema auf der ersten Achse. Anzumerken sei an dieser Stelle jedoch, daß die Verwendung von dichotomen Daten bei der Faktorenanalyse problematisch ist, da die Eingabedaten metrisches Skalenniveau haben müssen. Diese Annahme gilt zwar - wenn auch mit deutlichen Einschränkungen - für dichotome Daten (Mittelwerte können z.B. als prozentuale Anteile der beiden Variablenausprägungen interpretiert werden), doch kommt es insbesondere bei schiefen Verteilungen zu starken Verzerrungen der Ergebnisse (vgl. etwa Denz 1982).

Ein relativ häufiger Anwendungsfall der Faktorenanalyse ist in der Faktorialökologie zu finden, wo viele verschiedene Merkmale (Altersstruktur, Erwerbstätigkeit, Ethnien, Bildung, ...) in sozialen Räumen (i.d.R. Orts- oder Stadtteile) zu wenigen Faktoren zusammengefaßt werden. So findet Hamm (1979) bei seiner Analyse der Stadt Bern (164 statistische Quartiere, 64 Variablen) vier für ihn relevante Einheiten, die er anhand der auf den entsprechenden Achsen ladenden Variablen beschreibt. Den ersten Faktor, auf den insgesamt 14 Variablen laden, bezeichnet Hamm (1979, S. 193) als »Segregation der Unterschicht«. Auf diesen laden positiv u.a. die Variablen »Anteil von unteren Angestellten«, »Anteil von Beamten«, »Besuch einer höheren Schule« und »Anteil von Beschäftigten im tertiären Sektor«, während die Variablen »Anteil von Arbeitern«, »Anteil von un - und angelernten Arbeitern«, »Anteil von Ausländern« und »Anteil von Primarschulabschlüssen« negativ darauf laden. Rein formal bedeutet dies u.a., daß in den statistischen Quartieren von Bern, in denen der Anteil der Arbeiter hoch ist, der Anteil der unteren Angestellten und der Beamten

niedrig ist - und umgekehrt. Dieses, als formal zu bezeichnende, Ergebnis ist jedoch tautologisch und kann schon aus der Betrachtung der einfachen bivariaten Korrelationen bzw. anhand von Plausibilitätsüberlegungen abgeleitet werden: In den Gebieten, in denen eine Gruppe (z.B. Arbeiter) stark vertreten ist, muß eine dazu komplementäre (z.B. Angestellte oder Beamte) schwach vertreten sein, da sich die Anteile der einzelnen Gruppen (alle Variablen der Berufskategorisierung) sich zu hundert addieren. Doch ist dies mitnichten das eigentliche Ergebnis der Analyse. Inhaltlich relevant ist z.B., daß der »Anteil der Arbeiter« mit dem »Anteil der Ausländer« auf demselben Abschnitt derselben Achse lädt, beide Gruppen also in den gleichen Teilgebieten der Stadt relativ häufig bzw. relativ selten zu finden sind. Ferner lädt auf derselben Achse, aber auf der anderen Seite, also negativ mit den beiden Gruppen korreliert, die Variable »Anteil von Beamten«. Daraus läßt sich ableiten, daß diese Gruppe von Personen in anderen statistischen Quartieren ihre Präferenzen hat als Arbeiter und Ausländer. Laden zwei Variablen auf unterschiedlichen Achsen, so sind sie voneinander unabhängig. Hierfür ein kleines Beispiel: Hamms Ergebnissen zufolge lädt auf der zweiten Achse die Variable »Anteil der jungen Erwachsenen«, während auf der ersten Achse die Variable »Anteil der Arbeiter« lädt. Die jungen Erwachsenen sind also etwa gleich häufig in den Gebieten zu finden, in denen der Anteil der Arbeiter überdurchschnittlich hoch ist, als auch in den Gebieten, in denen der Anteil der Arbeiter unterdurchschnittlich ist.

Blotevogel (1979) verwendet die Hauptkomponentenanalyse zur Analyse der Wirtschaftstruktur deutscher Großstädte nach der Berufszählung von 1907. In einem ersten Schritt extrahiert Blotevogel aus 105 eingegebenen Berufsvariablen in 37 Städten sieben für ihn relevante Faktoren. Im zweiten Schritt ordnet er, mittels der Ausprägungen der Städte in den latenten Variablen, die 37 Städte den einzelnen Faktoren zu, um zu beschreiben, welche Städte ähnliche Strukturmerkmale aufweisen und welche sich voneinander unterscheiden.

Eine als konfirmatorische Faktorenanalyse zu bezeichnende Schätztechnik verwenden Falter et al. (1983) in ihrer Untersuchung zum Zusammenhang von Entstehungsbedingungen des Nationalssozialismus und Arbeitslosigkeit. Nachdem sie mögliche latente Faktoren und die darauf ladenden Variablen theoretisch hergeleitet haben, bestimmen sie die Korrelationen zwischen den latenten, sowie zwischen den latenten und manifesten Variablen. Als Ergebnis erhalten sie ein Pfaddiagramm (Falter et al. 1983, S. 548), welches sie anhand der direkten und indirekten Pfade - und den dazugehörigen Pfadkoeffizienten - interpretieren.

In den bisher genannten Anwendungsbeispielen wird die Faktorenanalyse primär als Verfahren zur Dimensionsanalyse verwendet, d.h. es wird getestet, welche Variablen eine gemeinsame Dimension bilden. Als

Faustregel läßt sich zu diesem Vorgehen angeben, daß mindestens drei bis vier Variablen auf einer Achse laden müssen, bevor im Sinne einer Dimensionsanalyse von »Dimension« gesprochen werden kann.

Primär als Datenreduktionsverfahren verwenden Blasius & Dangschat (1988) die Hauptkomponentenanalyse. Für die Erklärung von Segregation nach Bildungsgruppen, dargestellt am Fallbeispiel Warschau, verwenden sie Variablen wie »Anteile von Haushaltsgrößen« oder »Anteile von Altersgruppen«. Da sich die jeweiligen Anteile zu 100 Prozent addieren, können diese wegen ihrer linearen Abhängigkeiten (Multikollinearität) nicht zusammen in ein Pfadmodell einbezogen werden. Mittels der Hauptkomponentenanalyse werden jeweils mehrere manifeste Variablen (z.B. die prozentualen Anteile der verschiedenen Altersgruppen) zu einer oder zwei latenten zusammengefaßt, die dann als erklärende Variablen in das Modell einbezogen werden. Die inhaltliche Bedeutung, also welche Variablen eine gemeinsame Dimension bilden, ist bei dieser Untersuchung lediglich von sekundärer Bedeutung.

Wie aus den angeführten Beispielen schon ersichtlich wurde, gibt es eine Vielzahl von z.T. recht unterschiedlichen Möglichkeiten mittels der Faktorenanalyse Daten zu strukturieren und zu reduzieren. Da aber letztlich die meisten Verfahren auf die Hauptkomponentenanalyse zurückzuführen sind und diese auch der Regelfall bei der empirischen Anwendung ist (u.a. Voreinstellung bei SPSSX), wird dieser Ansatz schwerpunktmäßig vorgestellt.

## 3. Das Hauptkomponentenmodell

Ausgangspunkt ist ein Gleichungssystem, indem m - Variablen durch k - Faktoren ($k <= m$) reproduzierbar sind. Da alle Variablen standardisiert in die Analyse eingehen sollen, die absolute Höhe ihrer Ausprägungen somit unberücksichtigt bleibt, werden sie z - transformiert. Voraussetzung für eine derartige Transformation ist metrisches Meßniveau aller verwendeten Variablen, die zudem normalverteilt sein müssen. Das allgemeine Modell der Hauptkomponentenanalyse kann als lineares Gleichungssystem dargestellt werden:

$$z_1' = a_{11} f_1' + a_{12} f_2' + \cdots + a_{1k} f_k' + \cdots + a_{1m} f_m'$$
$$z_2' = a_{21} f_1' + a_{22} f_2' + \cdots + a_{2k} f_k' + \cdots + a_{2m} f_m'$$
$$\vdots$$
$$z_m' = a_{m1} f_1' + a_{m2} f_2' + \cdots + a_{mk} f_k' + \cdots + a_{mm} f_m'$$

Hierbei sind die $a_{11}$ bis $a_{mm}$ die (standardisierten) Regressionskoeffizienten der Variablen auf den Achsen. Sie werden in diesem Modell als Korrelationen interpretiert und üblicherweise als »Ladungen« bezeichnet. Die $z_1$ bis $z_m$ sind Zeilenvektoren mit n (= Anzahl der Untersuchungseinheiten, Stichprobenumfang) Komponenten. Die $f_1'$ bis $f_m'$ sind ebenfalls Zeilenvektoren mit je n Komponenten. Diese entsprechen den Ausprägungen der latenten, also den nicht - beobachtbaren Variablen der einzelnen Untersuchungseinheiten. Die Quadrate der $a_{11}$ bis $a_{mm}$ entsprechen der durch die jeweiligen Faktoren erklärten Varianz der Variablen, im Fall der vollständigen Lösung ist die Zeilensumme der $a_{.i}^2$ Eins. Diese aufsummierten erklärten Varianzen der Zeilen werden als Kommunalitäten bezeichnet. Hier ist abzulesen, wieviel Varianz der einzelnen Variablen durch die berücksichtigten k - Faktoren erklärt wird. Schreiben wir obiges Gleichungssystem in Matrixform, so ist:

(1) $Z = A \cdot F$

A wird üblicherweise als Ladungsmatrix und F als Faktorwertematrix bezeichnet. In diesem linearen Gleichungssystem kann es maximal soviele Faktoren geben, wie Variablen vorhanden sind; im Fall von k = m ist das Modell vollständig durch die errechneten Faktoren determiniert. Das steht allerdings im Gegensatz zu dem Ziel, mit möglichst wenigen Faktoren (k = minimal) die Daten optimal zu reproduzieren. Werden lediglich die ersten k - Faktoren berücksichtigt, so ist:

(2) $Z = A_k \cdot F_k + V$

wobei V der Fehlerterm ist, d.h. der Anteil der Variablen, der nicht durch die ersten k - Faktoren erklärt wird.

## 4. Die Korrelationsmatrix

Die Korrelation zwischen zwei z - transformierten Variablen kann formal als 1/n • z'z beschrieben werden, wobei n die Anzahl der Untersuchungseinheiten ist. Die Konstante 1/n entspricht dem mittleren Abweichungsquadrat des Skalarproduktes. (Als Skalarprodukt wird der Wert bezeichnet, der sich aus der Multiplikation eines Zeilenvektors mit einem Spaltenvektor ergibt, wobei beide Vektoren die gleiche Anzahl von Komponenten - z.B. Untersuchungseinheiten - haben müssen.) Die Korrelationen in einem Set von Variablen können dargestellt werden als:

**(3)** $R = 1/n \cdot Z'Z$

Da bei der Korrelation einer Variablen mit sich selbst immer Eins herauskommt, stehen in der Hauptdiagonalen der Korrelationsmatrix R nur Einsen. Bei m - Variablen ist die Spur (die Summe der Elemente in der Hauptdiagonalen) dieser Matrix m. Der (maximale) Rang einer Matrix (Anzahl der linear unabhängigen Zeilen/ Spalten) entspricht dem Minimum von Zeilen (Untersuchungseinheiten) oder Spalten (Variablen), wobei immer die Anzahl der Untersuchungseinheiten deutlich größer sein muß als die Anzahl der Variablen. Ist, wie z.B. bei Blotevogel (1979), die Anzahl der Untersuchungseinheiten kleiner als die Anzahl der Variablen, so kann es zwar rein rechnerisch soviele Faktoren wie Variablen geben, jedoch sind mindestens (m - n) Faktoren irrelevant. Auch inhaltlich kann diese Vorgehensweise zu keinen sinnvollen Ergebnissen führen, da Faktoren berechnet werden, die überhaupt keine Untersuchungseinheit (bei Blotevogel keine Stadt) beschreiben können. (Im Extremfall, also dort wo die Anzahl der Variablen gleich der Untersuchungseinheiten ist, kann eine latente Variable genau eine Untersuchungseinheit beschreiben, sie wäre praktisch das numerische Gegenstück, die exakte Beschreibung.)

Mit dem Hauptkomponentenmodell soll durch den ersten latenten Faktor ein Maximum der Gesamtvarianz erklärt werden. Der zweite soll von dem verbleibenden Rest der zu erklärenden Gesamtvarianz wiederum ein Maximum erklären, dabei aber unkorreliert (orthogonal) zu dem ersten Faktor sein. Der dritte Faktor soll ebenfalls wieder ein Maximum der verbliebenen Restvarianz erklären und orthogonal zu den beiden anderen Faktoren sein. Das gleiche gilt für alle nachfolgenden Faktoren, sie sollen jeweils ein Maximum der Restvarianz erklären und unkorreliert zu den jeweils anderen sein. Die Unkorreliertheit der Faktoren läßt sich grafisch als 90 - Grad Winkel (der Korrelationskoeffizient entspricht dem Kosinus des Winkels zwischen zwei Achsen, cos (90°) = 0) darstellen, in dem die einzelnen Achsen sich zueinander befinden. Durch diese Unkorreliertheit der Faktoren gilt im Fall der vollständigen Lösung:

(4) 1/n . F F « I,

wobei I die Einheitsmatrix ist. (In der Hauptdiagonalen stehen die Korrelationen der Faktoren mit sich selber, also lauter Einsen, in den übrigen Feldern die Korrelationen zwischen den Faktoren, die laut Modellanahme alle Null sind.) Da weniger Faktoren berücksichtigt werden sollen als zu einer vollständigen Lösung des Gleichungssystems notwendig sind, soll die Korrelationsmatrix (die ursprünglichen Variablen) mit den relevanten Faktoren optimal reproduziert werden.

## 5. **Die Kanonische Zerlegung**

Für jede symmetrische Matrix gilt, daß sie in die Matrix ihrer Eigenvektoren (E) und in eine Diagonalmatrix ihrer Eigenwerte ( $\Gamma$ ) wie folgt zerlegbar ist:

(5) $R = E \Gamma E'$

In der Diagonalmatrix der Eigenwerte sind die Komponenten, also die Eigenwerte zu den jeweiligen Achsen, der Größe nach geordnet. Die Summe der Eigenwerte ist gleich der Spur der Matrix R, bei einer Korrelationsmatrix somit gleich der Anzahl der Variablen. Damit diese Operation nachvollziehbar wird, haben wir die Korrelationsmatrix (Eingabedaten) sowie die Matrix der Eigenwerte und der Eigenvektoren angeführt.

Als Beispiel wählten wir eine Korrelationsmatrix, die wir der Studie von Heiland (1982, S.254) entnahmen. Heiland überprüfte mit einer Pfadanalyse, ob es einen Zusammenhang von Industrieproduktion (INDPRODK), Lebenshaltungskosten (LHKINDEX), dem durchschnittlichen jährlichen Einkommen aller Erwerbstätigen (JEINKERW), dem Anteil von Auswanderen und Arbeitslosen (AUS - ARBQ), dem privaten Verbrauch von Nahrungsmitteln (VBNAHRGM) und der Höhe der wegen einfachen Diebstahles festgenommen Personen (EINFDBST) gibt. Für seine Analysen verwendete Heiland trendbereinigte Zeitreihen (zur Notwendigkeit und der praktischen Ausführung dieser Bereinigung siehe z.B. Schlittgen & Streitberg 1987) aus den Jahren 1882 bis 1936, wobei er die Werte aus den Jahren des ersten Weltkrieges nicht berücksichtigte. Ziel seiner Arbeit ist es, den schon von Berg (1902) beschriebenen Zusammenhang von einfachem Diebstahl und Lebensmittelpreisen in einer etwas komplexeren Weise, vor allem aber mit einer statistisch ausgereifteren Methode zu überprüfen.

Wie aus den nachfolgenden Daten (siehe Daten 1) ersichtlich wird, existieren genausoviele Eigenwerte und Eigenvektoren wie Variablen. Die Summe der sechs Eigenwerte ergibt, wie bereits erläutert, die Spur der

DATEN 1

KORRELATIONSMATRIX DER EINGABEDATEN

| A | INDPRODK | LHKINDEX | JEINKERW | AUS-ARBQ | VBNAHRGH | EINFDBST |
|---|---|---|---|---|---|---|
| INDPRODK | 1.000 | -0.003 | 0.046 | -0.072 | 0.497 | -0.346 |
| LHKINDEX | -0.003 | 1.000 | -0.244 | 0.295 | -0.335 | 0.248 |
| JEINKERW | 0.046 | -0.244 | 1.000 | -0.397 | 0.423 | 0.118 |
| AUS-ARBQ | -0.072 | 0.295 | -0.397 | 1.000 | -0.310 | 0.480 |
| VBNAHRGH | 0.497 | -0.335 | 0.423 | -0.310 | 1.000 | 0.055 |
| EINFDBST | -0.346 | 0.248 | 0.118 | 0.480 | 0.055 | 1.000 |

EIGENVEKTOREN VON A

| E | 1.FAKTOR | 2.FAKTOR | 3.FAKTOR | 4.FAKTOR | 5.FAKTOR | 6.FAKTOR |
|---|---|---|---|---|---|---|
| ROW1 | 0.3331 | -0.2386 | 0.7316 | -0.0222 | -0.2654 | 0.4753 |
| ROW2 | -0.3963 | -0.0332 | 0.3521 | -0.8004 | 0.1427 | -0.2385 |
| ROW3 | 0.3928 | 0.5200 | -0.1748 | -0.3713 | -0.6278 | -0.1130 |
| ROW4 | -0.4904 | 0.1561 | 0.3832 | 0.4431 | -0.4724 | -0.4109 |
| ROW5 | 0.4921 | 0.3337 | 0.3856 | 0.1455 | 0.4802 | -0.4960 |
| ROW6 | -0.3082 | 0.7320 | 0.1211 | 0.0594 | 0.2478 | 0.5381 |

EIGENWERTE VON A

| N | COLI |
|---|---|
| 1.FAKTOR | 2.2182 |
| 2.FAKTOR | 1.3128 |
| 3.FAKTOR | 1.1545 |
| 4.FAKTOR | 0.7659 |
| 5.FAKTOR | 0.4061 |
| 6.FAKTOR | 0.1425 |

ERKLAERTE VARIANZ DER ACHSEN

| ERVAR | COLI |
|---|---|
| 1.FAKTOR | 0.3697 |
| 2.FAKTOR | 0.2188 |
| 3.FAKTOR | 0.1924 |
| 4.FAKTOR | 0.1277 |
| 5.FAKTOR | 0.0677 |
| 6.FAKTOR | 0.0237 |

*4*

Korrelationsmatrix, sie ist gleich der Anzahl der Variablen. Das Produkt der einzelnen Spaltenvektoren mit sich selbst ergibt in allen sechs Fällen Eins (So für den ersten Faktor: $0,3331^2 + -0,3963^2 + 0,3928^2 + -0,4904^2 + 0,4921^2 + -0,3082^2 = 1$), das Produkt zweier unterschiedlicher Spaltenvektoren immer Null. (So für den ersten und zweiten Faktor: $0,3331 \cdot -0,2386 + -0,3963 \cdot -0,0332 + 0,3928 \cdot 0,52 + -0,4904 \cdot 0,1561 + 0,4921 \cdot 0,3337 + -0,3082 \cdot 0,732 = 0$).

Die Werte der Eigenvektoren haben lediglich eine geometrische Bedeutung aber keinen inhaltlichen Bezug. Mittels der Eigenwerte kann der Anteil der einzelnen Achsen durch die erklärte Varianz bestimmt werden, um damit zu überprüfen, wieviele Dimensionen existieren. Hierfür werden die einzelnen Eigenwerte durch die Summe aller Eigenwerte dividiert. Für den ersten Eigenwert ergibt sich demzufolge ein Wert von 0,3697, d.h die dazugehörige erste Achse erklärt 36,97 Prozent der Gesamtvarianz; die zweite Achse erklärt weitere 21,88 Prozent. Wird noch die Varianz der dritten Achse hinzuaddiert, so erklären diese drei fast achtzig Prozent der Varianz der sechs manifesten Variablen. Eine inhaltliche Diskussion über die Anzahl der vorhanden Dimensionen soll an dieser Stelle jedoch unterbleiben, da hierfür zu wenig Variablen in die Analyse aufgenommen wurden.

## 6. **Bestimmung der Ladungsmatrix**

Wir hatten definiert, daß

(6) $Z = A \cdot F$ und daß

(7) $R = 1/n \cdot Z Z'$ ist.

Setzen wir in Gleichung (7) für Z die rechte Seite von Gleichung (6) ein, so erhalten wir:

(8) $R = 1/n \cdot A F F'A'$

Da $1/n \cdot F F = 1$ ist, folgt

(9) $R = A A'$

Gleichung (9) wird als Fundamentaltheorem der Faktorenanalyse bezeichnet. Da ebenso $R = E \Gamma E'$ gilt, ergibt sich für die Berechnung der Ladungsmatrix:

(10) $A = E \Gamma^{1/2}$

Um die Matrix der Korrelationen zwischen den latenten und den beobachteten Variablen (A, in der Darstellung als LADUNG bezeichnet) zu er-

halten, muß die Matrix der Eigenwerte radiziert und mit der Matrix der Eigenvektoren multipliziert werden. Im Fall der nicht - vollständigen Lösung werden die zu den nicht - berücksichtigten Faktoren dazugehörigen Eigenwerte auf Null gesetzt, d.h. in der Hauptdiagonalen der Matrix ($\Gamma$) stehen an den entsprechenden Stellen Nullen. Daraus folgt, daß die entsprechenden Spalten der Ladungsmatrix ebenfalls Null sind. Im folgenden ist die Matrix der Korrelationen zwischen manifesten und latenten Variablen für den Fall der vollständigen Lösung angegeben.

DATEN 2

LADUNGSMATRIX

| LADUNG | 1.FAKTOR | 2.FAKTOR | 3.FAKTOR | 4.FAKTOR | 5.FAKTOR | 6.FAKTOR |
|---|---|---|---|---|---|---|
| INDPRODK | 0.4961 | -0.2734 | 0.7861 | -0.0194 | -0.1691 | 0.1794 |
| LHKINDEX | -0.5903 | -0.0380 | 0.3783 | -0.7005 | 0.0910 | -0.0900 |
| JEINKERW | 0.5850 | 0.5959 | -0.1878 | -0.3249 | -0.4000 | -0.0426 |
| AUS-ARBQ | -0.7303 | 0.1789 | 0.4118 | 0.3878 | -0.3010 | -0.1551 |
| VBNAHRGH | 0.7330 | 0.3823 | 0.4143 | 0.1274 | 0.3060 | -0.1872 |
| EINFDBST | -0.4591 | 0.8387 | 0.1302 | 0.0520 | 0.1579 | 0.2031 |

Aus der Ladungsmatrix (LADUNG) (siehe Daten 2) wird ersichtlich, daß auf dem ersten Faktor alle Variablen laden. Es existiert ein positiver Zusammenhang zwischen der Industrieproduktion (INDPRODK), dem durchschnittlichen Jahreseinkommen aller Erwerbstätigen (JEINKERW) und dem Verbrauch an Lebensmitteln (VGNAHRGM). In den Jahren, in denen die Werte dieser Variablen hoch sind, sind niedrige Werte bei dem Lebenshaltungskostenindex (LHKINDEX), der Arbeitslosen - und Auswanderungsquote (AUS - ARBQ), sowie bei wegen einfachen Diebstahles verhafteten Personen (EINFDBST) zu verzeichnen. Während Heiland in seiner multivariaten Analyse keinen Zusammenhang zwischen Lebenshaltungskostenindex und Einkommen sowie zwischen dem Ausmaß der Industrieproduktion und der Arbeits - und Auswanderungsquote feststellte, sind beide Korrelationen unseren Ergebnissen zufolge negativ: In Jahren hoher Konjunktur ist die Arbeitslosen - und Auswanderungsquote niedrig und in den Jahren, in denen das jährliche Einkommen aller Erwerbstätigen hoch ist, ist der Lebenshaltungskostenindex niedrig - und umgedreht.

Ferner ermitteln wir im Gegensatz zu Heiland negative Zusammenhänge zwischen der Anzahl der wegen einfachen Diebstahles verhafteten Personen und dem Einkommen aller Erwerbstätigen sowie dem privaten Nahrungsmittelverbrauch, allerdings nur bezogen auf die erste Dimension.

Dementgegen existiert auf der zweiten Achse ein positiver Zusammenhang zwischen diesen drei Variablen, insbesondere zwischen dem jährlichen Einkommen und der Anzahl der wegen Diebstahles verhafteten Personen. Um zu prüfen, auf welcher Achse die Variablen nun tatsächlich laden, die nach der Hauptkomponentenanalyse mit unterschiedlichen Faktoren korreliert sind, kann eine Faktorenrotation (s. Kapitel 10) durchgeführt werden.

Wird das Fundamentaltheorem der Faktorenanalyse, R = A A\ lediglich auf zwei Variablen bezogen, so ist:

$$(11) \quad \sum_j a_{ij}\, a_{j1} = r_{i1}$$

also das Skalarprodukt zwischen den Ladungsvektoren $a_i$ und $a_1$ (summiert über j = 1 bis k), gleich dem Korrelationskoeffizienten zwischen den Variablen i und 1. Wurde das vollständige Faktorenmodell (k = m) gewählt, so entsprechen die zurückgerechneten Korrelationskoeffizienten den ursprünglichen.

Bei spaltenweiser Aufsummierung der quadrierten Ladungen (die quadrierten Korrelationen der Variablen mit den Faktoren) ergeben sich die Eigenwerte der Faktoren. Bei Zeilen weiser Aufsummierung der quadrierten Ladungen ergeben sich die Kommunalitäten der einzelnen Variablen, d.h. der Anteil der Varianz, der durch die berücksichtigten ersten k - Faktoren erklärt wird. Um dies anschaulich zu zeigen, wählen wir die Lösung, in der drei Faktoren berücksichtigt wurden.

Für den ersten Spaltenvektor der Ladungsmatrix »LADUNG« (siehe Daten 3) ergibt sich bei Summation der Quadrate seiner Komponenten der erste Eigenwert. (So ist $0{,}4961^2 + -0{,}5903^2 + 0{,}585^2 + -0{,}0303^2 + 0{,}733^2 + -0{,}4591^2 = 2{,}2182$.) Bei Aufsummierung der quadrierten Komponenten des ersten Zeilenvektors ergibt sich die erklärte Varianz (Kommunalität) der ersten Variablen (INDPRODK) unter Berücksichtigung der ersten drei Achsen. (So ist $0{,}4961^2 + -0{,}2734^2 + 0{,}7861^2 - 0{,}9388$) Anhand des Vektors der Kommunalitäten wird ersichtlich, daß die Variable »Lebenshaltungskostenindex (LHKINDEX)« lediglich zu 49,3 Prozent durch die ersten drei Achsen determiniert wird, während die Variable »Industrieproduktion« zu 93,88 Prozent erklärt wird.

Im nachfolgenden Abschnitt wird demonstriert, wie durch eine schrittweise Berücksichtigung der Faktoren sich die, aus dem Produkt $A_k \cdot A_k'$ zurückrechenbare, Korrelationsmatrix der ursprünglichen annähert. Bei Aufnahme aller m - Faktoren ist die Differenz aus der Eingabematrix und der zurückgerechneten Korrelationsmatrix die Nullmatrix. Es läßt sich somit zeigen, daß die Hauptkomponentenanalyse nichts weiter ist als eine durch die Faktoren erklärte Korrelationsmatrix zuzüglich einer Residualmatrix. Bei Berücksichtigung lediglich des ersten Faktors ergeben sich die

DATEN 3

LADUNGSNATRIX

| LADUNG | 1.FAKTOR | 2.FAKTOR | 3.FAKTOR | 4.FAKTOR | 5.FAKTOR | 6.FAKTOR |
|---|---|---|---|---|---|---|
| INDPRODK | 0.4961 | -0.2734 | 0.7861 | 0.0000 | 0.0000 | 0.0000 |
| LHKINDEX | -0.5903 | -0.0380 | 0.3783 | 0.0000 | 0.0000 | 0.0000 |
| JEINKERW | 0.5850 | 0.5959 | -0.1878 | 0.0000 | 0.0000 | 0.0000 |
| AUS-ARBQ | -0.7303 | 0.1789 | 0.4118 | 0.0000 | 0.0000 | 0.0000 |
| VBNAHRGM | 0.7330 | 0.3823 | 0.4143 | 0.0000 | 0.0000 | 0.0000 |
| EINFDBST | -0.4591 | 0.8387 | 0.1302 | 0.0000 | 0.0000 | 0.0000 |

KOMMUNALITAETEN

| KOMHU | COLI |
|---|---|
| INDPRODK | 0.9388 |
| LHKINDEX | 0.4930 |
| JEINKERW | 0.7326 |
| AUS-ARBQ | 0.7349 |
| VBNAHRGM | 0.8551 |
| EINFDBST | 0.9311 |

nachfolgenden 3 Matrixen und der Vektor der Kommunalitäten (siehe Daten 4).

Je höher die Variablen auf der ersten Achse laden, desto besser ist deren zurückgerechnter Wert. So ist nach diesem Schritt die Variable »Verbrauch von Nahrungsmitteln« schon zu 53,73 Prozent ($0{,}733^2 = 0{,}5373$) determiniert, wie sich an der Hauptdiagonale (5.Stelle) der Matrix »KOR« ablesen läßt. (Die Werte in der Hauptdiagonalen sind gleich den bereits erwähnten Kommunalitäten.) An dieser Stelle läßt sich auch leicht zeigen, daß die (zurückgerechneten) Korrelationen sich aus den Ladungen bestimmen lassen: So entspricht der Korrelationskoeffizient zwischen den Variablen INDPRODK und LHKINDEX dem Produkt der ersten beiden Zeilenvektoren, also 0,4961 • - 0,5903 + 0 . 0 + ... = - 0,2929. Wird die zurückgerechnete Korrelationsmatrix von der eingegebenen Korrelationsmatrix subtrahiert, ergibt sich die Differenzmatrix. Nach der Aufnahme des letzten Faktors ist die zurückgerechnete Korrelationsmatrix gleich der eingegebenen und die Differenzenmatrix die Nullmatrix. Von diesen weiteren Schritten soll jedoch nur noch die Aufnahme des zweiten Faktors dokumentiert werden. Bei Berücksichtigung der ersten beiden Faktoren verändern sich die in Daten 4 aufgeführten Matrixen und der Vektor der Kommunalitäten folgendermaßen (siehe Daten 5).

Mit diesen zwei Faktoren werden nun insgesamt 58,85 Prozent des Gesamtmodells erklärt. Durch diese zusätzliche Aufnahme steigt u.a. die er

DATEN 4

LADUNGSMATRIX

| LADUNG | 1.FAKTOR | 2.FAKTOR | 3.FAKTOR | 4.FAKTOR | 5.FAKTOR | 6.FAKTOR |
|---|---|---|---|---|---|---|
| INDPRODK | 0.4961 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| LHKINDEX | -0.5903 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| JEINKERW | 0.5850 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| AUS-ARBQ | -0.7303 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| VBNAHRGN | 0.7330 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| EINFDBST | -0.4591 | 0.0000 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

KOMMUNALITAETEN

| KOMMU | COLI |
|---|---|
| INDPRODK | 0.2462 |
| LHKINDEX | 0.3484 |
| JEINKERW | 0.3422 |
| AUS-ARBQ | 0.5334 |
| VBNAHRGM | 0.5373 |
| EINFDBST | 0.2108 |

ZURUECKGERECHNETE KORRELATIONSMATRIX

| KOR | INDPRODK | LHKINDEX | JEINKERW | AUS-ARBQ | VBNAHRGM | EINFDBST |
|---|---|---|---|---|---|---|
| INDPRODK | 0.2462 | -0.2929 | 0.2902 | -0.3623 | 0.3637 | -0.2278 |
| LHKINDEX | -0.2929 | 0.3484 | -0.3453 | 0.4311 | -0.4327 | 0.2710 |
| JEINKERW | 0.2902 | -0.3453 | 0.3422 | -0.4272 | 0.4288 | -0.2686 |
| AUS-ARBQ | -0.3623 | 0.4311 | -0.4272 | 0.5334 | -0.5353 | 0.3353 |
| VBNAHRGM | 0.3637 | -0.4327 | 0.4288 | -0.5353 | 0.5373 | -0.3365 |
| EINFDBST | -0.2278 | 0.2710 | -0.2686 | 0.3353 | -0.3365 | 0.2108 |

DIFFERENZ: AUSGANGSMATRIX-ERRECHNETER KORRELATIONSMATRIX

| DIFF | INDPRODK | LHKINDEX | JEINKERW | AUS-ARBQ | VBNAHRGM | EINFDBST |
|---|---|---|---|---|---|---|
| INDPRODK | 0.7538 | 0.2899 | -0.2442 | 0.2903 | 0.1333 | -0.1182 |
| LHKINDEX | 0.2899 | 0.6516 | 0.1013 | -0.1361 | 0.0977 | -0.0230 |
| JEINKERW | -0.2442 | 0.1013 | 0.6578 | 0.0302 | -0.0058 | 0.3866 |
| AUS-ARBQ | 0.2903 | -0.1361 | 0.0302 | 0.4666 | 0.2253 | 0.1447 |
| VBNAHRGM | 0.1333 | 0.0977 | -0.0058 | 0.2253 | 0.4627 | 0.3915 |
| EINFDBST | -0.1182 | -0.0230 | 0.3866 | 0.1447 | 0.3915 | 0.7892 |

***DATEH 5***

LADUNGSMATRIX

| LADUNG | 1.FAKTOR | 2.FAKTOR | 3.FAKTOR | 4.FAKTOR | 5.FAKTOR | 6.FAKTOR |
|---|---|---|---|---|---|---|
| INDPRODK | 0.4961 | -0.2734 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| LHKINDEX | -0.5903 | -0.0380 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| JEINKERW | 0.5850 | 0.5959 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| AUS-ARBQ | -0.7303 | 0.1789 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| VBNAHRGM | 0.7330 | 0.3823 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |
| EINFDBST | -0.4591 | 0.8387 | 0.0000 | 0.0000 | 0.0000 | 0.0000 |

KOMMUNALITAETEN

| KOMMU | COLI |
|---|---|
| INDPRODK | 0.3209 |
| LHKINDEX | 0.3499 |
| JEINKERW | 0.6973 |
| AUS-ARBQ | 0.5654 |
| VBNAHRGM | 0.6835 |
| EINFDBST | 0.9142 |

ZURUECKGERECHNETE KORRELATIONSMATRIX

| KOR | INDPRODK | LHKINDEX | JEINKERW | AUS-ARBQ | VBNAHRGM | EINFDBST |
|---|---|---|---|---|---|---|
| INDPRODK | 0.3209 | -0.2825 | 0.1274 | -0.4112 | 0.2591 | -0.4570 |
| LHKINDEX | -0.2825 | 0.3499 | -0.3680 | 0.4243 | -0.4472 | 0.2391 |
| JEINKERW | 0.1274 | -0.3680 | 0.6973 | -0.3207 | 0.6566 | 0.2312 |
| AUS-ARBQ | -0.4112 | 0.4243 | -0.3207 | 0.5654 | -0.4669 | 0.4853 |
| VBNAHRGM | 0.2591 | -0.4472 | 0.6566 | -0.4669 | 0.6835 | -0.0158 |
| EINFDBST | -0.4570 | 0.2391 | 0.2312 | 0.4853 | -0.0158 | 0.9142 |

DIFFERENZ: AUSGANGSMATRIX-ERRECHNETER KORRELATIONSMATRIX

| DIFF | INDPRODK | LHKINDEX | JEINKERW | AUS-ARBQ | VBNAHRGM | EINFDBST |
|---|---|---|---|---|---|---|
| INDPRODK | 0.6791 | 0.2795 | -0.0814 | 0.3392 | 0.2379 | 0.1110 |
| LHKINDEX | 0.2795 | 0.6501 | 0.1240 | -0.1293 | 0.1122 | 0.0089 |
| JEINKERW | -0.0814 | 0.1240 | 0.3027 | -0.0763 | -0.2-336 | -0.1132 |
| AUS-ARBQ | 0.3392 | -0.1293 | -0.0763 | 0.4346 | 0.1569 | -0.0053 |
| VBNAHRGM | 0.2379 | 0.1122 | -0.2336 | 0.1569 | 0.3165 | 0.0708 |
| EINFDBST | 0.1110 | 0.0089 | -0.1132 | -0.0053 | 0.0708 | 0.0858 |

klärte Varianz der sechsten Variablen (EINFDBST) von 21,08 Prozent auf 91,42 Prozent. Werden die Komponenten der Differenzenmatrixen (DIFF) in der sechsten Zeile/ Spalte nach dem ersten und dem zweiten Schritt miteinander verglichen, so fällt auf, daß sich deren absolute Beträge wesentlich verringert haben: Die Anpassung an die eingegebene Korrelationsmatrix ist deutlich besser geworden.

## 7. Bestimmung der Faktorenzahl

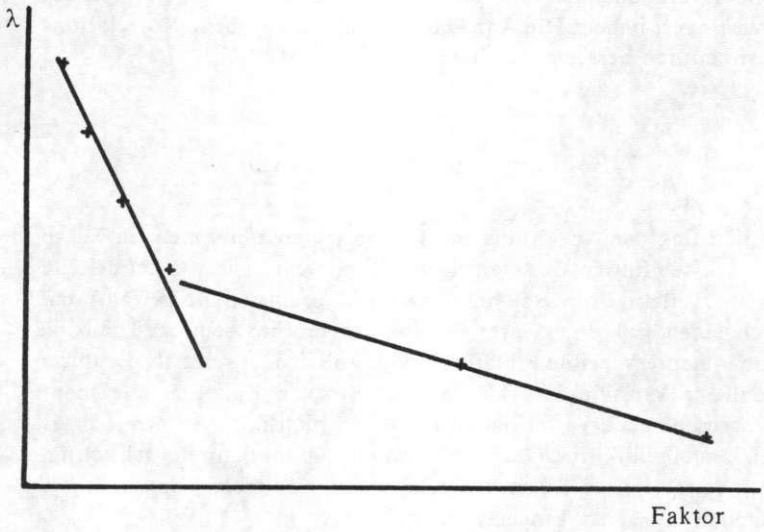
Die Bestimmung der Anzahl der in die Analyse aufzunehmenden Faktoren kann auf vielfältige Weise erfolgen, ein allgemein verwendetes Kriterium existiert nicht. So wurde bei SPSSX als Voreinstellung gewählt, daß die durch einen Faktor erklärte Varianz größer Eins sein muß, d.h die durch eine latente Variable erklärte Varianz soll größer sein, als die durch eine manifeste Variable zu erklärende. Dieses von Kaiser & Diekmann (1959) vorgeschlagene Verfahren ist gewiß ein plausibler Ansatz, doch ist er zugleich auch willkürlich: So wird etwa ein Faktor dann berücksichtigt, wenn der dazugehörige Eigenwert 1,0001 ist; ein anderer, dessen Eigenwert lediglich 0,9999 ist, hingegen nicht.

Eine andere Möglichkeit hat Catell (1966) vorgeschlagen. Seinem als Scree - Test bezeichneten Verfahren zufolge sollen soviele Faktoren aufgenommen werden, bis es einen Sprung im Eigenwertdiagramm (vgl. Abbildung) gibt. Doch ist auch dieses Vorgehen eher als Daumenregel zu bezeichnen, da dieses Kriterium oft nicht so eindeutig ist, wie es in der Abbildung erscheint.

Ein weiterer - ähnlich oft praktizierter - Vorschlag bezieht sich auf die Interpretierbarkeit der Faktoren. Soviele Faktoren wie (für den jeweiligen Forscher) interpretierbar sind, werden aufgenommen.

Neben diesen rein subjektiven Kriterien gibt es einen von Barlett (1951) entwickelten, auf der Chi - Quadrat - Verteilung basierenden Test, mit dessen Hilfe die signifikante Anzahl von Faktoren geschätzt werden kann. Jedoch ist dieser Test lediglich approximativ gültig (vgl. Ost 1984) und zudem abhängig vom Stichprobenumfang. Da uns ferner keine praktische Anwendung dieses Verfahrens bekannt ist, haben wir auf eine eingehende Darstellung verzichtet.

**Eigenwertdiagramm**




## 8. Weitere Probleme

Bei der Faktorenanalyse sind die Ladungsmatrix »LADUNG« und die Faktorwertematrix »F« empirisch kaum überprüfbar, da es sich den Modellanahmen zufolge um latente - also nicht (direkt) meßbare Variablen und deren Korrelationen handelt. Daher sind inhaltliche Ergebnisse oft nur schwer nachvollziehbar. Zudem sind die latenten Variablen quantitativ skaliert, d.h. es wird ein Kontinuum unterstellt, welches vielleicht gar nicht existieren kann.

Ein weiteres Problem ist die unterstellte Annahme, daß sich die manifesten Variablen als Linearkombination von latenten Variablen - zuzüglich eines vernachlässigbaren Fehlers - darstellen lassen. Daß diese Linearitätsannahme, die keineswegs nur für die Faktorenanalyse charakteristisch ist, zu verzerrten Ergebnissen führen kann, zeigt McDonald (1986). Er führte schon in den 60er Jahren (McDonald: 1967a, 1967b) ein faktorenanalytisches Modell ein, das nicht auf dieser Linearitätsannahme beruht - eine sozialwissenschaftliche Anwendung dieses Verfahrens ist uns jedoch nicht bekannt.

Da bei der Faktorenanalyse in der Regel Korrelationsmatrizen als Eingabeinformationen verwendet werden (es ist aber durchaus zulässig, auch Varianz - Kovarianzmatrizen einzulesen), fehlt das absolute Niveau der

Ausprägungen, was zur Folge hat, daß ein direkter Vergleich von Ergebnissen selten inhaltlich sinnvoll ist. Anders ausgedrückt: Die Faktorenanalyse ist bezugsgruppenabhängig. So können sich, z.B. bei Einstellungstests, Männer und Frauen hinsichtlich des absoluten Niveaus bei der Beantwortung von Fragen (Mittelwerte) erheblich unterscheiden, wobei die getrennten Interkorrelationsmatrizen - und damit auch die ermittelten latenten Faktoren - jedoch weitgehend identisch sind. Das gleiche Problem kann auch bei einer Messung über Zeit auftreten: Einstellungen können extremer werden, ohne daß dies eine Änderung der Faktorenstruktur zur Folge hat. Somit ist eine untersuchungsspezifische Objektivität der Messung nicht gewährleistet, die Faktorenanalyse für die Entwicklung von Meßskalen nur bedingt brauchbar.

## 9. Weitere Verfahren der Faktorenanalyse

Neben der Hauptkomponentenanalyse ist die von Thurstone (1947) entwickelte Hauptfaktorenanalyse in der empirischen Sozialforschung weit verbreitet. Im Unterschied zur Hauptkomponentenanalyse wird bei diesem Verfahren davon ausgegangen, daß neben den gemeinsamen Faktoren auch spezifische Faktoren existieren. Letztere entsprechen den Anteilen der Varianz, die nur für die einzelnen Variablen relevant sind, d.h. es existieren genausoviele spezifische Faktoren wie Variablen. Das Modell der Hauptfaktorenanalyse läßt sich darstellen als:

(12) $Z = A \cdot F + V,$

wobei V eine Matrix ist, in der die Anteile der einzelnen Variablen stehen, die nicht durch die gemeinsamen Faktoren erklärt werden. Für die Zerlegung der Korrelationsmatrix (R) bedeutet dies, daß nicht mehr diese zerlegt wird, sondern eine um die Fehlerterme reduzierte Korrelationsmatrix (R - U). In der Hauptdiagonalen von U steht der (geschätzte) Anteil der Varianz der m Variablen, der im Fall der vollständigen Lösung nicht durch die m gemeinsamen Faktoren erklärt wird. Die Matrix U ist eine Diagonalmatrix, in der die standardisierten Einzelvarianzen der Fehlerterme stehen, alle anderen Komponenten sind Null. (Von Null verschiedene Werte können in der vollständigen Lösung nicht stehen, da diese als gemeinsame Korrelationen von Variablen auf einer der Achsen laden würden.)

Für die Schätzung der Kommunalitäten bedeutet dies, daß die Zeilensumme der quadrierten Ladungen (Kommunalität der jeweiligen Variable) im Fall der vollständigen Lösung nicht mehr Eins ergibt, sondern daß von diesen Werten die Fehlervarianzen der Variablen subtrahiert werden müssen. Beim rechnerischen Vorgehen werden im ersten Schritt diese

Fehlervarianzen geschätzt (In der Regel wird in der ersten Schätzung der multiple Korrelationskoeffizient der einzelnen Variablen berechnet, dessen Quadrat gebildet und von Eins subtrahiert. Dieser Wert entspricht dem Anteil der nicht - erklärten Varianz der Variablen.) und die Differenz (R - U) bestimmt, anschließend wird diese reduzierte Matrix dann zerlegt, so daß:

(13) $R - U = E \Gamma E'$

Diese beiden Schritte der Hauptfaktorenanalyse lassen sich iterativ wiederholen bis eine zufriedenstellende Schätzung der Ladungsmatrix (LADUNG) und der Kommunalitäten bzw. der Residualmatrix (U) erreicht wurde (zum Schätzalgorythmus siehe z.B. Arminger, 1979; Ost, 1984). Die Matrix (R - U) ist zwar reell und symmetrisch, aber da die Anzahl der theoretisch errechenbaren Faktoren (k + m; im Fall der vollständigen Lösung ist k = m) größer als die Anzahl der Variablen ist, können negative Eigenwerte abgeleitet werden, was mathematisch wenig sinnvoll ist; die dazugehörigen Faktoren sind nicht interpretierbar. Da ferner die Kommunalitäten geschätzt und nicht errechnet werden, die Summe der Eigenwerte somit im Gegensatz zur Hauptkomponentenanalyse nicht analytisch bestimmt ist, von ihnen aber letztlich auch die Anzahl der Faktoren und deren geometrischen Lage im Raum abhängen, ist eine gewisse Willkürlichkeit bei diesem Verfahren vorhanden.

Die einzige, im eigentlichen Sinne statistische Methode der Faktorenanalyse, ist die von Lawley und Maxwell (1971) entwickelte ML - Faktorenanalyse. Mit Hilfe der Maximum - Likelihood - Methode werden die Parameter (insb. die Ladungsmatrix und der Vektor des Fehlerterms) derart bestimmt, daß die Stichprobe (oder genauer, die bei diesem Verfahren analysierte Kovarianzmatrix), aus der die genannten Parameter geschätzt werden, eine maximale Wahrscheinlichkeitsdichte aufweist. Da dieser Methode eine relativ komplexe Schätzstatistik zugrunde liegt und zudem bisher in der sozialwissenschaftlichen Forschung nur selten angewandt wurde, soll dieses Verfahren hier nicht explizit vorgestellt werden; eine einführende Darstellung geben z.B. Arminger (1979) und Ost (1984).

## 10. Faktorenrotation

Keine weiteren Verfahren der Faktorenanalyse, sondern lediglich Techniken zur Umstrukturierung der erhaltenen Faktoren sind Faktorrotationen, die in der Regel im Anschluß an eine Faktorenanalyse gemacht werden. Bereits in den vierziger Jahren stellte Thurstone (1947, S. 335) einen Forderungskatalog bezüglich der Einfachstruktur von Faktoren auf. Demzufolge besitzt eine Faktorlösung Einfachstruktur, wenn sie möglichst ein-

fache, d.h. theoretisch sparsame Erklärungen der Variablen mittels der Faktoren ermöglicht. Dies ist immer dann der Fall, wenn ein Teil der Ladungen Nahe Null (keine Korrelation) und der andere Teil der Ladungen Nahe Eins (hohe Korrelation) liegt. Mit Hilfe von Faktorrotationen wird versucht, dieser Forderung nahe zu kommen, oder anders ausgedrückt: Rotationen werden verwendet, um die Faktoren optimal mittels der Variablen interpretieren zu können. Bezogen auf unser Beispiel bedeutet dies, daß wir u.a. feststellen wollen, auf weichser Achse die Variablen »jährliches Einkommen aller Erwerbstätigen« und »wegen einfachen Diebstahls verurteilten Personen« laden.

Bei der Faktorenrotation wird das Ziel aufgegeben, daß der erste Faktor ein Maximum an Varianz erklären soll, der zweite dann wieder ein Maximum der verbleibenden Restvarianz usw., sondern die Faktoren sollen »optimal« die Variablenstruktur abbilden und möglichst einfach (eindeutig) interpretierbar sein. Der für die Identifikation der Achsen wichtige Indikator »Anteil der erklärten Varianz« wird dabei irrelevant. Im folgenden sollen kurz die wichtigsten Rotationsverfahren vorgestellt werden, für eine ausführliche Darstellung, insbesondere auch der algebraischen Herleitungen, siehe Arminger (1979), Holm (1976) oder Überla (1971).

Am verbreitetsten dürfte die von Kaiser (1958) vorgeschlagene VARIMAX - Rotation sein. Ausgehend von den errechneten Hauptkomponenten werden alle Achsen um einen (immer den gleichen) Winkel (a) gedreht, so daß die Varianz der quadrierten Ladungen auf den Achsen maximal wird, d.h. die Unterschiede in den quadrierten Ladungen auf den Faktoren sollen maximal sein. Im Falle einer optimalen Lösung wären alle Korrelationen zwischen den manifesten und den latenten Variablen nahe Null bzw. Eins, wobei die Orthogonalität der Faktoren erhalten bleibt. Im folgenden sind die nicht - quadrierten Ladungen auf den rotierten Faktoren angegeben. Für diese Rotation wurden die ersten drei Faktoren verwendet, da deren Eigenwerte größer Eins sind.

Nach dieser Rotation laden z.B. die Variablen INDPRODK und EINFDBST nicht mehr auf der ersten Achse, während die Variable JEINKERW nun ausschließlich auf dieser Achse lädt; entsprechend dem Kriterium vom Thurstone ist die Faktorenstruktur einfacher geworden (siehe Daten 6). Daß die erklärte Varianz der Variablen (Kommunalität) erhalten bleibt, kann leicht nachgerechnet werden, indem die Quadrate der Komponenten der Zeilenvektoren addiert werden. Entgegen der nicht - rotierten Lösung ergeben die spaltenweise Aufsummierungen der quadrierten Elemente aber nicht mehr die Eigenwerte.

Werden diese Ergebnisse inhaltlich interpretiert, so lassen sich auch hier - wie bei der unrotierten Lösung - die positiven Zusammenhänge von »Lebenshaltungskostenindex« und »Auswanderungs - bzw. Arbeitslosenquote« sowie vom »Einkommen aller Erwerbstätigen« und dem »Ver-

DATEN 6

LADUNGSMATRIX NACH VARIMAX-ROTATION

| | 1.FAKTOR | 2.FAKTOR | 3.FAKTOR |
|---|---|---|---|
| INDPRODK | -.06346 | -.16409 | .95283 |
| LHKINDEX | -.54410 | .44032 | .05525 |
| JEINKERW | .85319 | .05162 | .04440 |
| AUS-ARBQ | -.50793 | .69057 | -.00677 |
| VBNAHRGM | .64085 | .04877 | .66485 |
| EINFDBST | .19686 | .92318 | -.20019 |

brauch von Nahrungsmitteln« auf der ersten Achse nachweisen. Der positive Zusammenhang der beiden erstgenannten Variablen mit der Variable »Anzahl wegen einfachen Diebstahls verurteilte Personen« wird auf der zweiten Achse ersichtlich, der von »Industrieproduktion« und »Verbrauch von Nahrungsmitteln« auf der dritten Achse. Während es in der unrotierten Lösung sowohl einen positiven als auch einen negativen Zusammenhang der Variablen EINFDBST und JEINKERW gibt, sind nach der VARIMAX - Rotation beide Variablen unkorreliert; sie laden auf verschiedenen Achsen. Somit gibt es in dem Untersuchungszeitraum weder einen positiven noch einen negativen Zusammenhang zwischen dem jährlichen Einkommen aller Erwerbstätigen und der Anzahl der wegen einfachen Diebstahls verurteilten Personen; beide Variablen sind voneinander unabhängig.

Ein anderes Rotationsverfahren ist die schiefwinklige OBLIMIN - Rotation. Im Unterschied zur VARIMAX - Rotation wird hierbei die Unkorreliertheit der Faktoren aufgegeben, wodurch es zu Korrelationen zwischen den einzelnen latenten Variablen (Faktoren) kommt, die inhaltlich interpretiert werden können. Die Ladungen sind nach der OBLIMIN - Rotation aber nicht mehr als Korrelationen zwischen latenten und manifesten Variablen interpretierbar und auch im Fall der vollständigen Lösung ist die Summe der quadrierten Ladungen der einzelnen Variablen ungleich Eins. Bei dieser Art der faktorenanalytischen Anwendung werden in der Regel nicht erst die Faktoren errechnet, um sie anschließend schiefwinklig zu rotieren, sondern Berechnung und Rotation erfolgen simultan.

Am häufigsten benutzt für diese Art der konfirmatorischen Faktorenanalyse sind die LISREL (linear structurel - relations) - Modelle (Jöreskop 1967, 1969, 1970). Weniger bekannt, aber prinzipiell ähnlich, ist die von Lohmöller (1979, 1984) beschriebene und von Falter et al. (1983) ver-

wendete Methode der »latent variables path analysis with partial least - Square estimation« (LVPLS). Da LVPLS und LISREL, sowie einige andere hier nicht genannte Analysepakete wie RAM oder COSAN, sich letztendlich lediglich im Schätzverfahren bzw. bei der Modellbildung unterscheiden, soll darauf nicht weiter eingegangen werden, zur Differenzierung dieser Programme siehe z.B. Lohmöller (1988).

## 11. Alternative Verfahren

Wir haben einige Probleme bei der Anwendung und Interpretation der Faktorenanalyse dargestellt, im abschließenden Kapitel sollen zwei alternative Verfahren kurz vorgestellt werden. Das wohl größte Hindernis zur Anwendung der Faktorenanalyse ist, das - in der Regel benötigte - metrische Datenniveau. Gerade darüber wird aber in der empirischen Sozialforschung relativ selten verfügt.

Diese Einschränkung gilt nicht für das explorative Verfahren der Korrespondenzanalyse; diese Methode kann sogar bei relativ kleinem Stichprobenumfang zur multivariaten Analyse nahezu beliebig vieler (kategorialer) Variablen verwendet werden (siehe Blasius 1987, Blasius & Rohlinger 1988). Bei der Korrespondenzanalyse handelt es sich primär um ein Verfahren zur grafischen Darstellung von Zeilen und Spalten einer (mehrerer) zweidimensionalen Kontingenztabelle(n). Mit Zweidimensionalität ist gemeint, daß alle Kreuztabellen der »abhängigen« mit den »unabhängigen« Variablen (»abhängig« und »unabhängig« werden nicht im kausalen Sinne verstanden, sondern dienen zur Unterscheidung von Zeilen - und Spaltenvariablen) untereinander geschrieben werden. Dies hat den Vorteil, daß es bei der multivariaten Anwendung keine Probleme mit der Frequenz der Zellenbesetzung gibt, wie dies z.B. bei der log - linearen Analyse der Fall ist.

Wie bei der Hauptkomponentenanalyse gibt es einen Satz von orthogonal aufeinander stehenden Vektoren, die einen niederdimensionalen Raum aufspannen. Die inhaltliche Interpretation dieser Achsen erfolgt bei der Korrespondenzanalyse aber nicht mittels der darauf ladenden Variablen, sondern anhand der darauf ladenden Variablenausprägungen. Neben der Möglichkeit qualitative Daten multivariat auswerten zu können, liegt der größte Vorteil der Korrespondenzanalyse darin, daß als Eingabeinformationen Rohdaten verwendet werden können, es also nicht erst zu einer Umrechnung in Korrelationskoeffizienten kommt, wobei schon ein Teil der Informationen (z.B. das absolute Niveau) verloren gehen kann.

Ein weiteres alternatives Verfahren zur Faktorenanalyse ist die Multidimensionale Skalierung (MDS), die zur Beschreibung von Strukturen zwischen Variablen (oder Personen) verwendet wird. Als Eingabedaten

werden Ähnlichkeitsdaten (in der Regel Paarvergleichsdaten) verwendet, was eine spezielle Art des Untersuchungsdisgn erfordert. Dies soll an einem kleinen Beispiel - Ähnlichkeit von Berufen, nach Ansicht von Befragten - erläutert werden. Bei der Faktorenanalyse wird den Befragten eine Liste von Items vorgelegt anhand derer sie angeben sollen, inwieweit die einzelnen Berufe einem (mehreren) Merkmal(en) (z.B. »körperlich anstrengend«) entsprechen. Aus diesen Daten wird dann eine Korrelationsmatrix errechnet, die als Eingabeinformation für die Faktorenanalyse verwendet wird. Für die Multidimensionale Skalierung müssen die Befragten die einzelnen Berufe (jeweils zwei) in Bezug auf ihre Ähnlichkeit (Unähnlichkeit) hinsichtlich eines (mehrerer) Merkmale(s) miteinander vergleichen; diese Matrizen der Ähnlichkeiten der Berufe bilden dann das Eingabematerial. Da aber auch Korrelationskoeffizienten - wenn auch mit gewissen Einschränkungen - als Ähnlichkeitsmaß (Unähnlichkeitsmaß) zwischen zwei Variablen angesehen werden können, ist ein spezielle Art der Erhebung für die Multidimensionale Skalierung nicht unbedingt erforderlich.

Ähnlich wie bei der Faktorenanalyse und der Korrespondenzanalyse die erklärte Varianz der einzelnen Achsen ein gutes Maß für die Bedeutung der Faktoren ist, gibt es bei der MDS einen Stresswert, anhand dessen sich die Güte der (zweidimensionalen) Präsentation angeben läßt. Ebenso wie die Faktorenanalyse und die Korrespondenzanalyse ist die MDS ein Verfahren zur Datenreduktion, wobei im Unterschied zu den beiden anderen Analysetechniken die inhaltliche Bedeutung der Faktoren relativ unwichtig ist - entscheident ist hier die inhaltliche Nähe der Variablen. Je näher zwei Variablen beieinander liegen, desto ähnlicher sind sie.

Ein »Nachteil« der Multidimensionalen Skalierung besteht darin, daß die Distanzen zwischen den Variablen - ähnlich wie bei der Clusteranalyse - auf vielfache Art und Weise bestimmt werden können. Da erhaltene Ergebnisse stark von der gewählten Distanzmetrik abhängig sind, ist deren Wahl - und insbesondere deren Begründung - ein zentrales Problem bei der Anwendung der MDS. Dies ist ein Problem, was weder bei der Faktorenanalyse noch bei der Korrespondenanalyse existiert.

**References:**

ARMINGER, G. (1979), Faktorenanalyse. Stuttgart: Teubner.

BARTLETT, M.S. (1951), The Effect of Standardization of a Chi - Quadrat Approximation in Factor Analysis. Biometrika 38, 337 - 344.

BERG, H. (1902), Getreidepreise und Kriminalität in Deutschland seit 1882. In: Liszt, F. (Hrsg.), Abhandlungen des kriminalistischen Seminars, erster Band, 2.Heft, 275 - 323.

BEST, H. (1986), Struktur und Handeln parlamentarischer Führungsgruppen in Deutschland und Frankreich 1848/49. Köln: Habilitationsschrift an der Universität zu Köln.

BLASIUS, J. (1987), Korrespondenzanalyse - ein multivariates Verfahren zur Analyse qualitativer Daten. Historical Social Research - Historische Sozialforschung 42/43, 172 - 189.

BLASIUS, J. & ROHLINGER, H. (1988), Korrespondenzanalyse - ein multivariates Programm zur Auswertung von zweidimensionalen Kontingenztabellen. In: Faulbaum, F. & Uehlinger, H. - M. (Hrsg.), Fortschritte der Statistik - Software 1, 387 - 397. Stuttgart: Gustav Fischer.

BLASIUS, J. & DANGSCHAT, J. (1988), An Explanation of Residential Segregation by Education for One City. The Case of Warsaw. Eingereicht bei American Journal of Sociology.

BLOTEVOGEL, H.H. (1979), Faktorenanalytische Untersuchungen zur Wirtschaftsstruktur der deutschen Großstädte nach der Berufszählung 1907. In: Schröder, W.H. (Hrsg.), Moderne Stadtgeschichte, 74- 111. Stuttgart: Klett - Cotta.

CATELL, R.B. (1966), The Scree Test for the Number of Factors. Multivariate Behavioral Research 1, 245 - 276.

DENZ, H. (1982), Analyse latenter Strukturen. München: Francke.

FALTER, J.W.; LINK, A.; LOHMÖLLER, J. - B.; RIJKE, J.; SCHUMANN, S. (1983), Arbeitslosigkeit und Nationalsozialismus. Kölner Zeitschrift für Soziologie und Sozialpsychologie 35, 525 - 554.

GALTON, F. (1869), Hereditary Genius. London: Macmillan.

HAMM, B. (1979), Landnutzung und soziale Segregation. In: Hamm, B. (Hrsg.), Lebensraum Stadt, 181 - 200. Frankfurt/M.: Campus.

HEILAND, H. - G. (1982), Zum Einfluß sozio - ökonomischer Veränderung auf die Entwicklung der Kriminalitätsrate in den Jahren 1882 - 1936. Eine multivariate Reanalyse kriminalstatistischer Untersuchungen. In: Albrecht, G. & Brüsten, M. (Hrsg.), Soziale Probleme und soziale Kontrolle, 246 - 262. Opladen: Westdeutscher Verlag.

HOLM, K. (1976), Die Befragung 3. Die Faktorenanalyse. München: Frankke.

HOTELLING, H. (1933), Analysis of a Complex of Statistical Variables into Principal Components. The Journal of Educational Psychology 24, 417-441 und 498 - 520.

JÖRESKOP, K.G. (1967), Some Contributions to Maximum Likelihood Factor Analysis. Psychometrika 32, 443 - 482.

JÖRESKOP, K.G. (1969), A General Approach to Confirmatory Maximum Likelihood Factor Analysis. Psychometrika 34, 183 - 202.

JÖRESKOP, K.G. (1970), A General Method for the Analysis of Covariance Structures. Biometrika 57, 239 - 251.

KAISER, H.F. (1958), The Varimax Criterion for Analytic Rotation in Factor Analysis. Psychometrika 23, 187 - 200.

KAISER, H.F. *Sc* DIEKMANN, K.W. (1959), Analytic Determination of Common Factors. American Psychologist 14, 425.

LAWLEY, D.N. & MAXWELL, A.E. (1971)[1], Factor Analysis as a Statistical Method. London: Butterworths.

LOHMÖLLER, J. - B. (1979), Pfadanalyse mit latenten Variablen: Das Programm PLSC. München: Hochschule der Bundeswehr.

LOHMÖLLER, J. - B. (1984), LVPLS, Program Manual, Version 1.6. Köln: Zentralarchiv für empirische Sozialforschung.

LOHMÖLLER, J. - B. (1988), Die LV - Pfadanalyseprogramme PLS, LISREL, EQS, COSAN und RAM im Vergleich. In: Faulbaum, F. & Uehlinger, H. - M. (Hrsg.), Fortschritte der Statistik - Software 1, 54 - 64. Stuttgart: Gustav - Fischer.

MCDONALD, R.P. (1967a), Numerical Methods for Polynominal Models in Nonlinear Factor Analysis. Psychometrika 32, 77 - 112.

MCDONALD, R.P. (1967b), Nonlinear Factor Analysis. Psychometric Monograph No. 15, 32.

MCDONALD, R.P. (1986), Describing the Elephant: Structure and Function in Multivariate Data. Psychometrika 51, 513 - 534.

MuLAIK, S.A. (1986), Factor Analysis and Psychometrika: Major Developments. Psychometrika 51, 23 - 33.

OST, F. (1984), Faktorenanalyse. In: Fahrmeir, L. & Hamerle, A. (Hrsg.), Multivariate statistische Verfahren, 575 - 662. Berlin: de Gruyter.

PEARSON, K. (1901), On Lines and Planes of Closest Fit to a System of Points in Space. Philosophical Magazine 2, 6th series, 557 - 572.

SCHLITTGEN, R. & STREIBERG, B. (1987)[2], Zeitreihenanalyse. München: Oldenbourg.

SPEARMAN, C. (1904), General Intelligence Objectively Determined and Measured. American Journal of Psychology 15, 201 - 293.

THURSTONE, L.L. (1947), Multiple - Factor Analysis. Chicago: University of Chicago Press.

ÜBERLA, K. (1971), Faktorenanalyse. Berlin: Springer.